Perspective-taking with Robots: Experiments and models

J. Gregory Trafton
Naval Research Laboratory
Code 5515
Washington, DC 20375-5337
trafton@itd.nrl.navy.mil

Alan C. Schultz
Naval Research Laboratory
Code 5515
Washington, DC 20375-5337
schultz@aic.nrl.navy.mil

Magdalena Bugajska
Naval Research Laboratory
Code 5515
Washington, DC 20375-5337
magda@aic.nrl.navy.mil

Farilee Mintz
Naval Research Laboratory
ITT Industries
mintz@aic.nrl.navy.mil

***Abstract*— We suggest that to enable effective human-robot interaction, robots should be able to interact in a way that is natural to and preferred by humans. Using human-compatible representations and reasoning mechanisms should help in developing skills which support effective human-robot interaction. In this paper, we present two studies that examine a critical human-robot-interaction component: perspective-taking. We find that when a person asks a robot to perform a task with some ambiguity to the robot, the person prefers the robot to either ask for clarification or take the person's perspective and act appropriately.**

I. Introduction

Imagine a robot collaborating with a person: the person is in charge, while the robot is the helper. The robot may be in charge of providing different tools to the person and generally helping out [1]. How should this collaboration proceed? We know that people use a great deal of spatial perspective-taking when working with another person; the speaker taking another person's point of view, or forcing the listener to take their own point of view [2]–[4]. We analyzed a large corpus of data from a training mission of astronauts collaborating on an assembly task; the brief discussion in Table I is typical of the types of conversations that the astronauts make while working together.

TABLE I

Example of Astronauts working together. They are working to unfasten some cables that are somewhat tangled and trying to determine the best order for removing them.

Bob	John	Perspective
	What do you recommend for taking them all off?	
Let's start with J78		object-centered
	Uh-huh	
Which is the one that's most forward [of you]		addressee-centered
	Uh-huh, most forward	egocentric

Notice that the first astronaut, John, suggests a plan of action (starting from a specific cable) and takes the perspective of the second astronaut. The second astronaut, Bob, then acknowledges the plan from his own point of view. An in-depth analysis of 4000 utterances of these astronauts showed that approximately 25% of the time perspective-taking was needed [4]. From this type of data, we suggest that if robots will be able to collaborate with people in these types of tasks, they must be able to not only understand their own perspective but also be able to take another person's perspective.

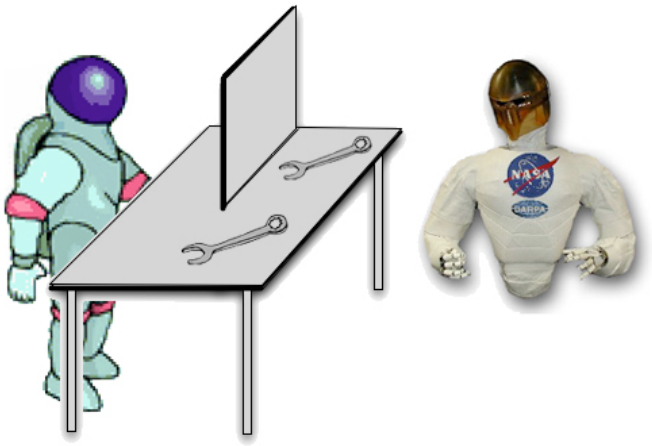

Fig. 1. When told "Give me the wrench," the robot needs to take the perspective of the person to determine which wrench the astronaut has referred to.

Clearly, the ability to take another person's perspective could be extremely useful, especially when there is ambiguity. For example, Figure 1 shows a situation where perspective-taking could be immediately useful. The robot and the person are facing each other; the robot can see that there are two wrenches on the table, but the astronaut only sees one wrench because the second wrench is occluded by an obstacle. When the astronaut says, "Robot, give me the wrench," the meaning of the phrase "the wrench" is ambiguous for the robot because it knows of two wrenches. The phrase is unambiguous to the astronaut, because he only knows about one wrench.

Other researchers have suggested that when people are in this type of situation they use the principles of least effort (people get things done in the easiest possible manner) and joint salience (people use what is salient to each other given the context) [5] and therefore, they would immediately reach for the wrench that both robot and person could see. If the robot could take the perspective of the astronaut, it would see that only one wrench is in the astronaut's field of view

and could therefore also surmise that "the wrench" must refer to the wrencth that both speaker and listener could see. Even in this rudimentary scenario, perspective-taking would immediately enhance the human-robot interaction.

It could be, however, that people would not trust a robot to deal with such ambiguities; that they would prefer the robot to explicitly ask any time there is ambiguity of this type. Taken to extremes, this "Always ask" policy would become extremely irritating, but one would expect that with the increase in complexity of the task, even humans would need to ask for additional assistance.

In the remainder of this paper, we present an experiment that explores people's preferences toward a robot taking their own perspective. In addition, we present a discussion of a system description of our own robot that is able to take the perspective of others.

II. Experiment

Our goal in this experiment was to explore the situation where a robot could take the perspective of another person, and then act on that knowledge.

A. Method

In this experiment, we were primarily interested in preference rather than raw performance; if people were not comfortable with a robot being able to take their perspective and then acting on that information, performance would be a irrelevant issue. Toward this end, we filmed different scenarios between a human (the speaker) and a robot (the listener) and asked participants to rank-order the scenarios by preference.

A second issue we explore in this experiment is whether peoples' preference changes if they observe the scenario from a listener's point of view or from a speaker's point of view.

We used an isomorph of the "Give me the wrench" scenario above: since our robots do not have hands or manipulators, we had the person ask the robot to go to a large traffic cone in the room.

1) Participants: Twenty-four participants from the Naval Research Laboratory were asked to make decisions about which scenario they preferred. Twelve participants were placed in the "Speaker (human) perspective" condition and twelve participants were placed in the "Listener (robot) perspective" condition. Order was counterbalanced and randomized by block: all possible orders were presented to the same number of participants to minimize order effects.

2) Materials: All scenarios were filmed in our robot laboratory. For all conditions, the person (speaker) could see only one traffic cone, while the robot (listener) could see two different cones (similar to the "wrench" example described earlier). For both perspective conditions, the materials consisted of an orienting scenario and three different "action" scenarios. The orienting scenario for both conditions showed the person's location and the stated command ("Go to the cone"). The three different action scenario conditions presented different ways that the robot could go to the cone. In the "Ask" scenario, the robot asked "Which cone?" and in response the person pointed and gestured to the only cone he could see, and the robot went to that cone. In the "Visible" scenario, the robot went to the cone that both the robot and the person could see. In the "Hidden" scenario, the robot went to the cone that only the robot could see. Participants were shown the different action scenario films in block-randomized order. An example of the display shown to the participants is shown in Figure 2.

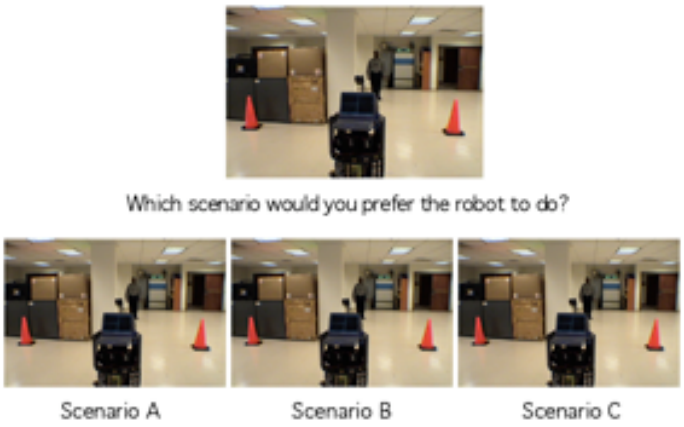

Fig. 2. Display as shown to the participants. Top movie shows the orienting scenario, and the bottom three movies show possible action scenarios that the subject will rank, "Ask," "Visible," and "Hidden" in random order.

The only difference between perspective conditions was the perspective from which the the interaction was shown. For the listener perspective condition, all movies were from the listener's perspective (e.g., from the robot's perspective). Participants viewed the scenarios from a camera located just behind the robot, as seen in Figure 3(b). For the speaker perspective condition, all movies were from the speaker's perspective (e.g., from the human's perspective). Participants in this condition viewed the scenario from a camera positioned behind the human, as in Figure 3(a). So while in both conditions, the listener (robot) could see two cones and the speaker (human) could only see one, the participants' view of the cones dependent on the perspective condition they were shown.

3) Procedure: Participants sat at a computer desk. They were told that they were going to see a person giving a command to the robot and then the robot would perform the action in one of three possible ways. After viewing the orienting and action scenarios, the participant rank ordered the different scenarios with 1 being the most preferred and 3 being the least preferred, based on the question, "Which scenario would you prefer the robot to do?". Participants could see any of the scenarios as many times as they liked.

After finishing the task participants were debriefed.

B. Results and Discussion

Because this experiment used rankings, non-parametric statistical tests were used. The results are summarized in Figure 4.

First, we examined the differences between perspective conditions. There were no differences of rank orderings between

(a) Speaker's perspective

(b) Listener's perspective

Fig. 3. Experimental setup for the two different perspective conditions.

the listener and speaker, $\chi^2(2) = 2.29, p > 0.3$.

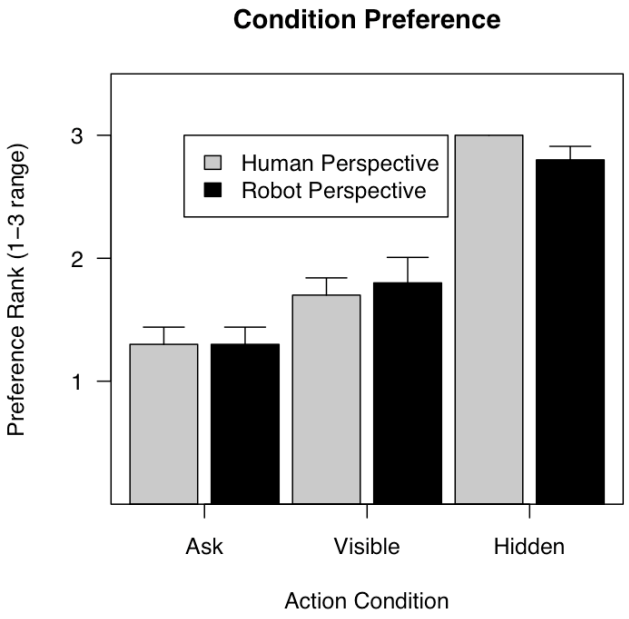

Fig. 4. Mean rankings by condition. Each bar represents the average ranking for that condition. A rank of 1 is the most preferred; a rank of 3 is the least preferred. The error bars represent standard error of the mean.

Second, we examined the differences between the different action scenarios. An omnibus Friedman's test showed that there was a significant difference between action scenarios from the perspective of both the speaker (human) $\chi^2(2) = 14, p < 0.001$ and the listener (robot), $\chi^2(2) = 18.7, p < 0.001$. However, this result only shows that there is a difference somewhere within the action conditions; it does not show which conditions are different from each other. Therefore, we performed a post-hoc comparison between action conditions using the procedure discussed in [6], which is a Bonferroni-type adjustment for multiple comparisons. Results of this analysis showed that there were no statistical differences between the Ask and Visible conditions, all $p > 0.2$, while the Hidden condition was statistically different for both the speaker, $p < 0.05$ and listener, $p < 0.05$ conditions. Thus, participants thought that when there was some ambiguity in the robot's response, it was acceptable to either ask the person for clarification or to simply take the person's perspective and act appropriately. Going to a hidden cone that the person had no knowledge of was clearly inappropriate.

It is evident, however, from looking at Figure 4 that the Ask condition is a bit preferred than the Visible condition. Even though we had enough statistical power to show that the hidden condition was a poor choice, it could be that we simply did not run enough participants (i.e., we did not have enough statistical power) to find differences between the Ask and the Visible condition. Since participants in the two perspective conditions (listener and speaker) did not differ in their responses, we combined them into one dataset (N=24) and re-ran the analysis. The omnibus Friedman test again showed a difference between action conditions, $\chi^2(2) = 32.3, p < 0.001$, and the post-hoc comparison showed that the Ask and Visible conditions were again statistically indistinguishable, while the Hidden conditions differed from both conditions, $p < 0.05$. Thus, it seems that the small numeric difference between Ask and Visible is merely a slight preference rather than an overwhelming lack of trust on the robot's part. Since we believe that people will ask for assistance under complex situations and/or situations with extreme ambiguity, it is clear that this scenario is not complex enough or too ambiguous to force people to ask for assistance or prefer the robot to confirm the choice.

C. Experiment Discussion

Participants appreciated that under ambiguous situations, a robot helper could either ask for assistance or take the person's perspective and act accordingly. This finding held true from both the speaker's and listener's perspective.

The remainder of the paper will discuss a robotic system that takes another person's spatial perspective when there is ambiguity.

III. Spatial Perspective Taking on a Robot

It is clear that if humans are to work as peers with robots in shared space, the robot must be able to understand the

natural human tendency to switch perspectives and to use different frames of reference. To create robots with these capabilities, we develop computational cognitive models of skills such as perspective-taking, and then use them as reasoning mechanisms for the robot. This approach has several benefits. First, a natural and intuitive interaction results in reduced cognitive load. Second, more predictable behavior engenders trust. Finally, more understandable decisions allow the human to recognize and more quickly repair mistakes in the interaction. [7]

A. Computational Cognitive Model of Perspective Taking

In our latest work, we used Polyscheme [8] to implement a computational cognitive model of perspective-taking. Polyscheme is a cognitive architecture aims to model how humans integrate multiple methods of representation, reasoning, and problem solving. Polyscheme has previously been integrated into a robotic architecture to provide symbolic reasoning and planning algorithms while maintaining the flexibility and robustness of reactive control systems [7], [9].

Given the position of the human and the robot, the positions of objects in the environment including the one to which the speaker referred to, the model, using mental simulations of the environment from different perspectives, is able to resolve ambiguity.

B. Perspective-Taking Task Examples

Using this model we have demonstrated a robot being able to solve several related perspective-taking tasks [4]. Videos of a robot and human in these tasks can be seen at "http://www.nrl.navy.mil/aic/iss/aas/CognitiveRobots.php".

In the first task, the robot is asked to "Go to the Cone" in a set up that is identical to the human subject experiment described earlier in Section II-A. Following the principle of least effort and joint salience described earlier, the perspective-taking model will allow the robot to move to the traffic cone that both the human and the robot can see.

In the second task, we demonstrate the generality of the perspective-taking model. In this task, there is only a single traffic cone that the human can see, but which is occluded from the robots perspective. The human asks the robot to "Go to the cone." The robot looks around and cannot detect a cone from its perspective. However, the robot can determine that there is a space that the human can see which it cannot see, and so, using the model, determines that the traffic cone must be in that area, and proceeds to look for the cone behind that occluded space.

Notice that we assume that the human is benevolent and not trying to fool the robot. We are only concerned at this stage of the research in the pure spatial perspective taking and not additional factors that may influence the robots behavior.

C. Robot Implementation

Whereas the computational cognitive model performs the high-level perspective taking, the actual mobility of the robot is handled by a navigation, localization and mapping system [10]. After the cognitive model determines where the robot should move, the mobility system handles the navigation to that location without collision.

The robot perception is handled with a map of the environment that the robot builds on the fly and with color tracking. The map is used for localization, path planning, navigation and collision avoidance, and gets its data from sonar and a structured light range finder. The color tracking is used to detect the traffic cones and to identify the occluding book shelves in the environment, and its data comes from an inexpensive web camera.

The robot is able to handle multi-modal interactions, using a combination of speech and gestures [11], although for these scenarios, only a simple utterances is required.

The robot platform itself is a Nomadic Technology Nomad 200, which has a three-wheeled synchronous steering base and a separately steerable turret with the sensors.

IV. Conclusion

We believe that to enable effective human-robot interaction, the robots should be able to interact in a way that is natural to the humans. Using similar representations and reasoning mechanisms should help in developing skills such as discussed perspective-taking, that are important to effective human-robot interaction.

Evidence suggests that humans use the principles of joint salience and least effort, as suggested in [5], in order to disambiguate and solve scenarios such as those presented here. In this paper, we presented the results of an empirical study that shows that in similar situations, humans will prefer a robot to be able to take the person's perspective to solve the task.

The results presented here, as well as the prior implementation of a model on a robot, are a first step towards robust, natural human-robot interaction. Much work still remains. Our models are of pure perspective taking. We assume the speaker is benevolent and that there are no tricks. We do not consider under what circumstances the task becomes complex enough that using these principles is sufficient to avoid asking for clarification. For example, the wrenches may be different in some way that would affect the task at hand, and the listener clearly should ask if he feels that the speaker should know about the unseen wrench. In these cases we would also want the robot to ask.

However, it is clear from our results that in many cases, not asking is the right and expected choice, and we have implemented a computational cognitive model that allows a robot to reason from another person's (or robot's) perspective. This skill is an important one for the robot to collaborate naturally with human teammates.

Acknowledgment

This work was supported by the Office of Naval Research under work order number N0001402WX20374. Additional funding was supplied by DARPA IPTO under the MARS program. The authors would like to thank William Adams, Brandon James, and Joseph Blumenthal for their help with

the materials for the experiment. The views and conclusions contained in this document should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U. S. Navy.

REFERENCES

[1] W. Bluethmann, R. Ambrose, M. D. andS. Askew, E. Huber, M. Goza, F. Rehnmark, C. Lovchik, and D. Magruder, "Robonaut: A robot designed to work with humans in space," *Autonomous Robots*, vol. 14, no. 2-3, pp. 179–197, 2003.

[2] N. Franklin, B. Tversky, and V. Coon, "Switching points of view in spatial mental models," *Memory and Cognition*, vol. 20, no. 5, pp. 507–518, 1992.

[3] H. A. Taylor and B. Tversky, "Perspective in spatial descriptions," *Journal of Memory and Language*, vol. 35, no. 3, pp. 371–391, 1996.

[4] J. G. Trafton, N. L. Cassimatis, M. D. Bugajska, D. P. Brock, F. E. Mintz, and A. C. Schultz, "Enabling effective human-robot interaction using perspective-taking in robots," *IEEE Trans. Syst., Man, Cybern. A*, in press.

[5] H. H. Clark, *Using language*. New York, NJ: Cambridge University Press, 1996.

[6] S. Siegel and N. J. Castellan, *Nonparametric statistics for the behavioral sciences*, 2nd ed. Boston, MA: McGraw-Hill, 1988.

[7] J. G. Trafton, A. C. Schultz, N. L. Cassimatis, L. M. Hiatt, D. Perzanowski, D. P. Brock, M. D. Bugajska, and W. Adams, "Cognition and multi-agent interaction: From cognitive modeling to social simulation," in *Communicating and collaborating with robotic agents*, R. Sun, Ed., in press.

[8] N. L. Cassimatis, "A cognitive architecture for integrating multiple representation and inference schemes," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, 2002.

[9] N. L. Cassimatis, J. G. Trafton, M. D. Bugajska, and A. C. Schultz, "Integrating cognition, perception and action through mental simulation in robots," *Journal of Robotics and Autonomous Systems*, vol. 49, no. 1-2, pp. 12–23, 2004.

[10] A. C. Schultz, W. Adams, and B. Yamauchi, "Integrating exploration, localization, navigation and planning through a common representation," *Autonomous Robots*, vol. 6, no. 3, pp. 293–308, 1999.

[11] D. Perzanowski, A. C. Schultz, W. Adams, E. Marsh, and M. Bugajska, "Building a multimodal human-interface," *IEEE Intelligent Systems and Their Applications*, vol. 16, no. 1, 2001.